Towards Semantic Integration for Explainable Artificial Intelligence in the Biomedical Domain

Catia Pesquita
LASIGE, Faculdade de Ciências da Universidade de Lisboa, Portugal

Keywords: Ontology Alignment, Knowledge Graph Alignment, Ontology Matching, Knowledge Graphs, Ontologies, Semantic Web, Explainable Artificial Intelligence, Machine Learning, Healthcare, Clinical Research, Health Informatics.

Abstract: Explainable artificial intelligence typically focuses on data-based explanations, lacking the semantic context needed to produce human-centric explanations. This is especially relevant in healthcare and life sciences where the heterogeneity in both data sources and user expertise, and the underlying complexity of the domain and applications poses serious challenges. The Semantic Web represents an unparalleled opportunity in this area: it provides large amounts of freely available data in the form of Knowledge Graphs, which link data to ontologies, and can thus act as background knowledge for building explanations closer to human conceptualizations. In particular, knowledge graphs support the computation of semantic similarity between objects, providing an understanding of why certain objects are considered similar or different. This is a basic aspect of explainability and is at the core of many machine learning applications. However, when data covers multiple domains, it may be necessary to integrate different ontologies to cover the full semantic landscape of the underlying data. We propose a methodology for semantic explanations in the biomedical domain that is based on the semantic annotation and integration of heterogenous data into a common semantic landscape that supports semantic similarity assessments. This methodology builds upon state of the art semantic web technologies and produces post-hoc explanations that are independent of the machine learning method employed.

1 INTRODUCTION

Recent successes in black-box models, such as deep neural networks, are revolutionizing artificial intelligence (AI) applications, but despite their impressive successes, their effectiveness and integration in real-world applications are still limited by their inability to explain their decisions in a human-understandable way. These limitations stem from ethical concerns, but also accountability, safety and liability (Guidotti et al., 2018). In critical use cases, for instance, in clinical decision making, there is reluctance in the deployment of such models because the cost of misclassification is potentially very high, endangering patients' health and lives (Miotto et al., 2018). Moreover, models that predict natural phenomena may better contribute to scientific advancements when researchers are able to understand them. This is evident in the application of black-box models in genomics, drug-discovery and pathology, among others (Min et al., 2017).

This is one of the historical challenges of AI: the ability of a model to afford explanations of how and why it arrived at a particular outcome. However, the definition of explainable AI (XAI) is still not agreed upon by the community and is often used interchangeably with interpretable or comprehensible AI (Guidotti et al., 2018). While interpretation requires transparency in the underlying mechanisms of a system, a comprehensible one can be opaque while emitting symbols a user can reason over. Both enable explanations of decisions, but they do not yield explanations themselves, leaving explanation generation to human analysts who may deduce different explanations depending on their background knowledge about the data and its domain. Relatively few works address these issues and they are typically based on researchers' intuitions of what constitutes a 'good' explanation, without taking into account how humans explain decisions and behaviour to each other, arguably a strong starting point to improve human interactions with explanatory AI (Miller, 2019).

Pesquita, C.
Towards Semantic Integration for Explainable Artificial Intelligence in the Biomedical Domain.
DOI: 10.5220/0010389707470753
In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021) - Volume 5: HEALTHINF*, pages 747-753
ISBN: 978-989-758-490-9

2 SEMANTIC TECHNOLOGIES FOR XAI IN BIOMEDICINE

One can argue that what is needed for humans to understand each other can be transferred to what is needed to make AI outcomes understandable for humans. We need to fulfil the properties of human understanding, namely, that human explanations imply social interaction which is grounded in a shared context, and that users select explanations from a large space of possible explanations based on their understanding of the context (Miller, 2019). The vast majority of works in XAI lack the ability to integrate background knowledge into the process to create a shared context, rendering them inadequate to build explanations for common users without AI expertise. Providing this contextualization is an even bigger challenge in areas such as systems medicine where data comes from different domains and with different levels of granularity or personalized medicine, which often relies on highly diverse data, ranging from molecules, organelles, cells, tissues, organs, all the way up to individuals, environmental factors, populations, and ecosystems (Holzinger et al., 2019).

2.1 Explainable Knowledge-enabled Systems

In the scientific and healthcare domains, where machine learning (ML) methods methods and particularly black-box methods such as deep learning have been gaining traction, it has been proposed that successful explainable-AI systems need to be able to link ML models to representations of domain knowledge (Holzinger et al., 2017; Wollschlaeger et al., 2020). Recently (Chari et al., 2020) defined Explainable Knowledge-enabled systems as "AI systems that include a representation of the domain knowledge in the field of application, have mechanisms to incorporate the users' context, are interpretable, and host explanation facilities that generate user-comprehensible, context-aware, and provenance enabled explanations of the mechanistic functioning of the AI system and the knowledge used."

However, most existing approaches that tackle XAI in knowledge-enabled systems focus on interpretability and not on building explanations. A recent survey (Chari et al., 2020) in this area presents neuro-symbolic approaches (Hitzler et al., 2020) as a potential solution, and although some preliminary works tackle explanation of deep learning for image recognition (e.g., (Zhou et al., 2018; Sarker et al., 2017)) or transfer learning (Chen et al., 2018) most simply allow for the inclusion of knowledge in the machine learning approaches but do not yield explanations themselves. Few XAI approaches exist in clinical application areas and most are still focused in statistical explanations (e.g.,(Lundberg et al., 2018)). (Phan et al., 2017) employs neuro-symbolic learning to predict human behaviour, producing simple explanations based on the identification of key features. However, explanations are flat, limited in expressivity.

2.2 Ontologies and Knowledge Graphs

We have established that for AI outcomes to be truly useful, they need to support interpretation, and this requires semantic context. By semantic context, we mean the situation in which a term or entity appears. In relational databases and spreadsheets, semantic context is sometimes lacking because important information about what the various data fields mean and how they relate to one another is often implicit in the names of database tables and column headers. What is needed is a way to express the semantic connections between data items in a way that is expressive enough to capture nuanced relationships while at the same time formalized and restrictive enough to allow software as well as humans to make inferences based on the links.

Semantic web (SW) technologies and artefacts (such as ontologies and knowledge graphs) are a potential solution to the problem of human-centric, knowledge-enabled explanation since they provide this semantic context (Lecue, 2019; d'Amato, 2020). Ontologies establish a conceptual model that represents the concepts of a domain and their relationships with one another (Munn and Smith, 2013), in a way that can be understood both by humans and machines. In general, an ontology encoding a domain of knowledge reflects the consensus of specialists or communities dealing with that domain. Domains are expressed mainly through classes referencing real-world entities (e.g. "Head", "Fever"), and the relationships found between them (e.g. "Face is part of Head", "Fever is a Vital Sign Finding"). The classes are frequently accompanied by lexical properties, such as preferred labels, synonyms, textual descriptions, etc., which provide human-readable definitions. In contrast, relationships between classes are represented not in text but through formal axioms, or statements, which make their intended meaning more amenable to automatic manipulation and reasoning, thus allowing the use of modern computational power to operate on the meaning, rather than the structure, of the real-world entities. A knowledge graph (KG) is frequently taken to mean a collection of data items with relations es-

tablished between them and described according to an ontology. The ontological layer of a KG thus describes and imposes some order on the data of a domain of interest.

In the biomedical and healthcare domains, the Semantic Web represents an unparalleled opportunity since it provides large amounts of freely available data and a set of technologies dedicated to data sharing, integration, management, and reasoning (Ferreira et al., 2020). The availability of over 1,000 open biomedical ontologies in BioPortal and more than 2 billion data items publicly available as a KG (i.e., Linked Open Data) represents a unique opportunity to integrate clinical and biomedical data. In the biomedical domain, these datasets range from semantic annotations for the functions of gene products, abnormal phenotypes related to diseases and genes, or drug adverse events. In electronic health records, patient-level data is commonly described using standardized coding schemes and ontologies, such as UMLS, SNOMED-CT and ICD-9/10. This annotation is mostly confined to final diagnosis and procedures, which are frequently used for billing purposes, whereas finer-grained clinical information is typically found in free text format, making its linking to ontologies a greater challenge. However, once data and AI outcomes are integrated with ontologies and KGs, they can serve as background knowledge to XAI applications and in this way afford the semantic context that is essential for explanations closer to human conceptualizations and thus more useful in real-world applications.

3 A METHODOLOGY FOR SEMANTIC EXPLANATIONS

XAI techniques can be categorized according to how they support human reasoning into inductive reasoning, querying and similarity modelling (Wang et al., 2019). Semantic explanations can actually support the three kinds of explanations, since KGs naturally support reasoning, querying and semantic similarity computation. To produce semantic explanations we need to address three challenges: (1) how to link input data and AI outcomes to their meaning; (2) how to link this meaningful data with what is already known and (3) how to use this contextual information to build effective explanations.

Figure 1 depicts the proposed methodology for semantic explanations in biomedical AI applications. It builds upon our well-established experience in semantic technologies within the biomedical domain and addresses the main challenges faced when building semantic explanations.

The core of the methodology is an integrated KG that supports the XAI approaches. The integrated KG is built by connecting heterogeneous data (scientific, clinical, etc.) to existing domain ontologies to provide a rich semantic layer to the data. This is achieved by performing **(1) Ontology selection** to determine the optimal set of ontologies to adequately describe the data, **(2) Semantic annotation**, to link the data to the ontologies, and **(3) Semantic integration** to establish links between the ontologies. The final step is to build **(4) Semantic explanation** approaches that explore background knowledge afforded by the KG. This methodology focuses on model agnostic explanations which are able to work regardless of the machine learning model employed and can be integrated into already existing approaches.

Let us consider a simple example of semantic explanations in healthcare, where we have trained a machine learning model that identifies patients with respiratory tract infections based on EHR data. Three patients arrive at the hospital with similar complaints. Patient A is described as having ”fever”, B as having ”fever w/ infection/cough”, and C as having a ”respiratory tract infection”. Figure 2 describes the semantic annotation of these cases using a subgraph of the SNOMED-CT.

Our machine learning model has classified both B and C as positive examples. While for patient C the classification as having a respiratory tract infection is straightforward, the same is not true for patient B. However, by having the patients described within the same semantic landscape which now includes background knowledge about symptoms and diseases, we are able to understand why patient B was also classified as having a respiratory tract infection. Notice that although the semantic annotation links each record to different concepts in the ontology, it is possible to reason that both patients B and C are more similar to each other than to patient A since both B and C suffer from an ”Infectious process” associated with a ”Respiratory finding”. Of course in this simplified example, patients are only described by easily comprehensible textual features, but the methodology extends to cases where objects are described my several annotations, across multiple domains.

The following describes each step of the methodology, highlights the main challenges that need to be addressed and provides a brief overview of state of the art semantic web technologies and tools that can be employed to tackle them.

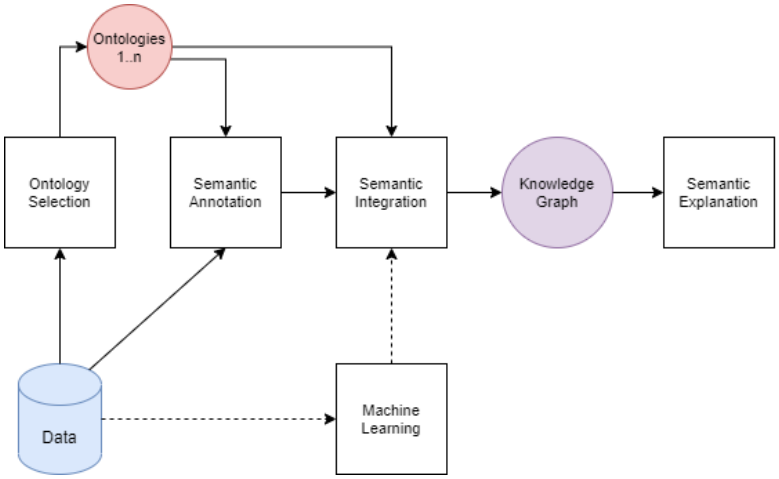

Figure 1: A methodology for semantic explanations for heterogeneous biomedical data.

3.1 Ontology Selection

Regarding the first challenge, it is expectable that multiple ontologies are needed to achieve a good coverage of semantic annotations, especially in multi-domain applications. However, to improve reasoning support, the selection should be focused on the minimum set of ontologies that still provides adequate granularity and scope to ensure the best possible coverage. Automated ontology recommendation services, such as BioPortal Recommender are able to recommend one or more ontologies that provide adequate coverage for input textual data (Martínez-Romero et al., 2017) considering aspects such as complementarity of resources and semantic richness. When multiple ontologies are selected, they should as much as possible be aligned and integrated to build a single unified semantic landscape. In previous work, we have developed automated approaches to select appropriate ontologies for data integration follwoing these principles (Faria et al., 2014).

3.2 Semantic Annotation

Once suitable ontologies are selected, the next step is semantic annotation, i.e. connecting data to its meaning that is encoded in an ontology creating a knowledge graph. To ensure a high-quality semantic description of the data explored by the machine learning models, we need to not only annotate feature values but also metadata (e.g. feature labels) and classification targets in the case of supervised learning. The semantic annotation of biomedical text needs to address several challenges (Jovanović and Bagheri, 2017): in the case of clinical notes, the use of abbreviation and acronyms as well as the prevalence of spelling mistakes and of meaningless notes (e.g., filling in a mandatory field with a period); in the case of biomedical terms, there is a high degree of ambiguity both in terms of polysemy and homonymy, which is further compounded with the use of acronyms that correspond to words (e.g. the CAT gene). These challenges are addressed by semantic annotation tools specifically designed for the semantic annotation of biomedical and clinical text (Tchechmedjiev et al., 2018) and recent advances in word embeddings specifically trained in the biological (Lee et al., 2020) and clinical domains (Alsentzer et al., 2019) are improving the performance in biomedical semantic annotation (Gonçalves et al., 2019).

3.3 Semantic Integration

It is not uncommon that a specific application requires multiple ontologies to describe the underlying data. One one hand, because complex applications such as clinical care and research require the integration of multiple domains of knowledge, and on the other because uncoordinated development often results in the adoption of multiple ontologies and controlled vocabularies that cover the same or similar domains. In these situations, where more than one ontology is used to annotate the data and to establish a ”shared context”, we need to identify the connections and relations between different ontologies and knowledge graphs. Discovering the semantic links or alignments between ontologies and the data sets that they organize can be very difficult, particularly if the datasets are large and complex, as is routinely the case in the biomedical domain. Biomedical and clinical datasets are particularly challenging to align for several reasons. Massive amounts of multimodal

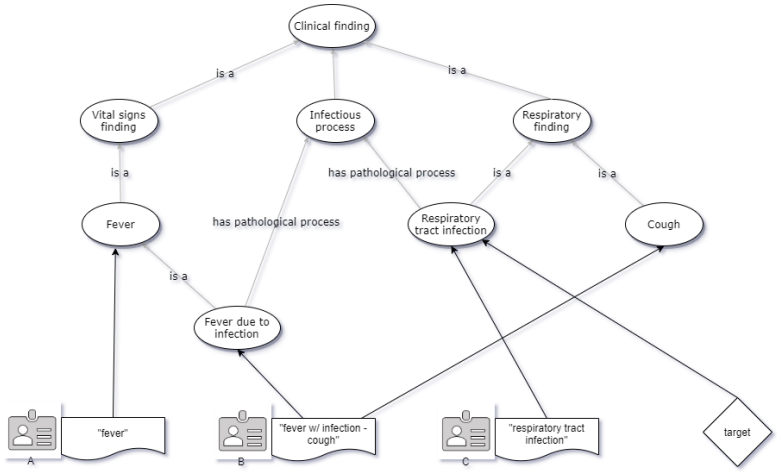

Figure 2: Example of the semantic annotation of patient records and machine learning target.

and diverse data are currently being generated by researchers, hospitals and mobile devices around the world, and their combined analysis presents unique opportunities for healthcare, science, and society. The data can range from molecular to phenotypic, behavioural to clinical, individual to population, genetic to environmental. Biomedical Big Data goes well beyond the recognized challenges in handling large volumes of data or large numbers of data sources, and presents specific challenges pertaining to the heterogeneity and complexity of data as well as to the complexity of its subsequent analysis.

Recent advances in semantic technologies support the rapid integration of datasets automatically or semi-automatically by encoding machine-readable representations of the meanings of data and metadata items; in particular, the success of the Linked Data movement demonstrates both the possibility of, and need for, semantic integration of data from diverse sources on a massive scale. This challenge can be addressed by employing ontology matching and linked data matching techniques which are able to identify meaningful links between entities described with different ontologies or vocabularies, in effect building a KG that connects all relevant entities through contextualized relations.

However, finding these relations is challenging, because biomedical vocabulary is rich and complex, different ontologies may model related concepts in a different way, and the relations between concepts may be themselves complex and semantically rich. In previous work, we have developed the AgreementMakerLight ontology matching system which employs a set of diverse computational techniques ranging from lexical matching, to machine learning, reasoning, graph visualization and user interaction to perform the alignment of ontologies and knowledge graphs. It is particularly suited to handle the challenges in biomedical ontology matching (Faria et al., 2018), including finding complex correspondences (Oliveira and Pesquita, 2018).

3.4 Semantic Explanations

When the outcomes of the ML models are semantically integrated with the input data, a shared semantic landscape can be explored by explanations. Although explanations based on reasoning and querying can be employed with the proposed methodology, since KGs naturally support both activities, here we focus on semantic similarity-based explanations. A natural process in human learning is to identify similar and distinguishing features to group similar objects and discriminate different ones. At their core, many types of AI approaches take into account similarity modelling, including distance-based methods, such as clustering; classification into different kinds, such as supervised learning; and dimensionality reduction, such as matrix factorization or autoencoders. Explanations for these approaches can be based on understanding why certain objects are considered similar or different (Wang et al., 2019).

The semantic similarity between two objects can be measured by comparing the ontology entities that describe them(Pesquita, 2017). A simple semantic similarity measure based on the ratio of shared ontology classes, would score the similarity between patient B and patient C as 3/6. Computing the similarity between the classification target and each patient is also possible, with patient A having a score of 1/6, B having a score of 1, and patient C having a score of 2/5. Both types of similarities, between instances, and between an instance and a classification target can be presented as explanations. There is

a large number of semantic similarity measures that take into account different ontology properties and object properties, providing more sophisticated similarity measures(). There are several challenges in measuring semantic similarity in the biomedical domain, namely how to address the multiple aspects that a KG can represent in the context of a specific application (Sousa et al., 2020), how to adequately consider the specificity of ontology classes (Aouicha and Taieb, 2016) and how to employ multiple ontologies(Ferreira and Couto, 2019). We have extensive experience in biomedical semantic similarity, having developed methods for its computation and evaluation, e.g. (Pesquita, 2017; Cardoso et al., 2020).

4 CONCLUSIONS

This work proposes a methodology to enable semantic explanations of machine learning applications in the biomedical domain. The methodology tackles the main challenges in providing human-centric explanations based on a contextualized understanding of the data and AI outcomes. It leverages the large amounts of freely available biomedical data and metadata in the form of Knowledge Graphs, and builds upon state of the art solutions for semantic annotation and integration to embed the data and AI outcomes with already established knowledge within the domain. It then explores semantic similarity between instances and between instances and outcomes to support similarity-based explanations. This methodology affords post-hoc explanations that are built independently of the machine learning algorithms employed, and can thus be integrated into any application for which data can be semantically annotated with existing biomedical ontologies.

In future work, we will employ this methodology to build a semantic explanation system integrating our existing contributions in semantic annotation, integration and similarity and apply it to the explanation of biomedical machine learning applications, including protein-protein interaction prediction, gene-disease association and disease progression prediction.

ACKNOWLEDGEMENTS

This work was funded by the Portuguese FCT through the LASIGE Research Unit (UIDB/00408/2020 and UIDP/00408/2020), and also by the SMILAX project (PTDC/EEI-ESS/4633/2014).

REFERENCES

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. (2019). Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Aouicha, M. B. and Taieb, M. A. H. (2016). Computing semantic similarity between biomedical concepts using new information content approach. *Journal of biomedical informatics*, 59:258–275.

Cardoso, C., Sousa, R. T., Köhler, S., and Pesquita, C. (2020). A collection of benchmark data sets for knowledge graph-based similarity in the biomedical domain. *Database*, 2020.

Chari, S., Gruen, D. M., Seneviratne, O., and McGuinness, D. L. (2020). Directions for explainable knowledge-enabled systems.

Chen, J., Lecue, F., Pan, J., Horrocks, I., and Chen, H. (2018). Knowledge-based transfer learning explanation. In *16th International Conference on Principles of Knowledge Representation and Reasoning*, pages 349–358. AAAI Press.

d'Amato, C. (2020). Machine learning for the semantic web: Lessons learnt and next research directions. *Semantic Web*, (Preprint):1–9.

Faria, D., Pesquita, C., Mott, I., Martins, C., Couto, F. M., and Cruz, I. F. (2018). Tackling the challenges of matching biomedical ontologies. *Journal of biomedical semantics*, 9(1):4.

Faria, D., Pesquita, C., Santos, E., Cruz, I. F., and Couto, F. M. (2014). Automatic background knowledge selection for matching biomedical ontologies. *PloS one*, 9(11):e111226.

Ferreira, J. D. and Couto, F. M. (2019). Multi-domain semantic similarity in biomedical research. *BMC bioinformatics*, 20(10):23–31.

Ferreira, J. D., Teixeira, D. C., and Pesquita, C. (2020). Biomedical ontologies: Coverage, access and use. In Wolkenhauer, O., editor, *Systems Medicine Integrative, Qualitative and Computational Approaches*, pages 382 – 395. Academic Press, Elsevier.

Gonçalves, R. S., Kamdar, M. R., and Musen, M. A. (2019). Aligning biomedical metadata with ontologies using clustering and embeddings. In *European Semantic Web Conference*, pages 146–161. Springer.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Hitzler, P., Bianchi, F., Ebrahimi, M., and Sarker, M. K. (2020). Neural-symbolic integration and the semantic web. *Semantic Web*, 11(1):3–11.

Holzinger, A., Biemann, C., Pattichis, C. S., and Kell, D. B. (2017). What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., and Müller, H. (2019). Causability and explainability of

artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4):e1312.

Jovanović, J. and Bagheri, E. (2017). Semantic annotation in biomedicine: the current landscape. *Journal of biomedical semantics*, 8(1):44.

Lecue, F. (2019). On the role of knowledge graphs in explainable ai. *Semantic Web*, (Preprint):1–11.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K.-W., Newman, S.-F., Kim, J., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature biomedical engineering*, 2(10):749–760.

Martínez-Romero, M., Jonquet, C., O'connor, M. J., Graybeal, J., Pazos, A., and Musen, M. A. (2017). Ncbo ontology recommender 2.0: an enhanced approach for biomedical ontology recommendation. *Journal of biomedical semantics*, 8(1):21.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38.

Min, S., Lee, B., and Yoon, S. (2017). Deep learning in bioinformatics. *Briefings in bioinformatics*, 18(5):851–869.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246.

Munn, K. and Smith, B. (2013). *Applied ontology: An introduction*, volume 9. Walter de Gruyter.

Oliveira, D. and Pesquita, C. (2018). Improving the interoperability of biomedical ontologies with compound alignments. *Journal of biomedical semantics*, 9(1):1.

Pesquita, C. (2017). Semantic similarity in the gene ontology. In *The gene ontology handbook*, pages 161–173. Humana Press, New York, NY.

Phan, N., Dou, D., Wang, H., Kil, D., and Piniewski, B. (2017). Ontology-based deep learning for human behavior prediction with explanations in health social networks. *Information sciences*, 384:298–313.

Sarker, M. K., Xie, N., Doran, D., Raymer, M., and Hitzler, P. (2017). Explaining trained neural networks with semantic web technologies: First steps. In *Twelfth International Workshop on Neural-Symbolic Learning and Reasoning 2017, London, UK, July 17-18, 2017.*

Sousa, R. T., Silva, S., and Pesquita, C. (2020). Evolving knowledge graph similarity for supervised learning in complex biomedical domains. *BMC bioinformatics*, 21(1):6.

Tchechmedjiev, A., Abdaoui, A., Emonet, V., Melzi, S., Jonnagaddala, J., and Jonquet, C. (2018). Enhanced functionalities for annotating and indexing clinical text with the ncbo annotator+. *Bioinformatics*, 34(11):1962–1965.

Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. (2019). Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15.

Wollschlaeger, B., Eichenberg, E., and Kabitzsch, K. (2020). Explain yourself: A semantic annotation framework to facilitate tagging of semantic information in health smart homes. In *HEALTHINF*, pages 133–144.

Zhou, B., Bau, D., Oliva, A., and Torralba, A. (2018). Interpreting deep visual representations via network dissection. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2131–2145.